DOCUMENT RESUME

ED 480 065 CG 032 638

AUTHOR	McDivitt, Patrica Jo
TITLE	Training Educators To Develop Good Educational Tests.
PUB DATE	2003-08-00
NOTE	17p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Accountability; *Educational Assessment; *Educational Testing; Learning Motivation; Standards; *Test Construction; Test Interpretation; *Training

ABSTRACT

In today's educational setting, assessment results weigh heavily in determining what students should know and be able to do. In addition, because assessment scores are often tied to accountability systems that affect both teaching and learning, they influence what is taught in the classroom. The changes in the use of assessments underscore the need for educators to understand the role assessment plays in instruction and learning. Teachers, counselors, and assessment professional are challenged not only to understand the use of assessments and the interpretation of the results, but also to learn more about how assessments are developed. This chapter summarizes the importance of assessment training with an emphasis on the standards-based assessment development process as it relates to recent research in learning and motivation. (Contains 12 references.) (GCP)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 065

Training Educators to Develop Good Educational Tests

By
Patricia Jo McDivitt

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

CG032638

Chapter 30
Training Educators to Develop Good Educational Tests

Patricia Jo McDivitt

In today's educational setting, assessment results weigh heavily in determining what students should know and be able to do. In addition, because assessment scores are often tied to accountability systems that affect both teaching and learning, they influence what is taught in the classroom. "Assessment directly affects learning in that it provides the necessary feedback for effective learning. It indirectly affects learning in that instruction is commonly skewed toward what is assessed; and, obviously, what is taught affects what is learned" (Marzano, Pickering, & McTighe, 1993, p. 11). The changes in the use of assessments underscore the need for educators to understand the role assessment plays in instruction and learning. Teachers, counselors, and assessment professionals are challenged not only to understand the use of assessments and the interpretation of the results, but also to learn more about how assessments are developed. This chapter summarizes the importance of assessment training with an emphasis on the standards-based assessment development process as it relates to recent research in learning and motivation.

Assessment Training Model

Although there are many models for assessment training, this chapter discusses in detail one suggested model. The components of this model are meant to enhance educators' understanding of the link between assessment and classroom teaching and learning. These components reflect recommended steps in the assessment development process:

- defining the purpose of the assessment
- developing content domains and understanding validity
- developing an assessment blueprint
- developing assessment question specifications
- writing and reviewing assessment questions

Defining the Purpose of the Assessment

The first step in assessment development training is to help educators understand the purpose of a particular assessment. Only by fully understanding this purpose can educators use the test scores appropriately. "In the broadest sense, the purpose of an assessment is to gather data to facilitate decision making. But there are many kinds of decisions and many kinds of information that may facilitate such decisions" (Mehrens, 2000, p. 27). Knowing precisely what students will be asked to master is important because different achievement targets require the application of different assessment methods, and there is no single assessment method capable of assessing all the various forms of achievement (Stiggins, 1999). Understanding the purpose of the standardized assessments used in the classroom is a fundamental component of any comprehensive assessment training program. This section summarizes the purpose of the two most common standardized assessments used in today's schools: the norm-referenced assessment and the criterion-referenced, or standards-based, assessment.

Norm-Referenced Assessment

The purpose of many norm-referenced standardized achievement assessments is to measure the academic foundation skills that students need. Therefore, the assessment questions are usually designed to measure a generalized set of objectives that are common across the country for a given content area. The results of this type of assessment allow educators to compare performance of students and to determine relative strengths and weaknesses of students based on the generalized academic foundation skills being measured by the assessment.

Norm-referenced standardized tests are based on national samples of students as the norm group for interpreting relative standing. Because these tests are designed for use in different schools throughout the country, they tend to provide broad coverage of each content area to maximize potential usefulness in as many schools as possible. Thus, educators should closely inspect the objectives and question types to determine how well the test matches the emphasis in the local curriculum (McMillan, 1997).

Criterion-Referenced, or Standards-Based, Assessment

Criterion-referenced assessments typically measure students' mastery of the learning targets as defined by a set of specific curriculum content standards. The use of standards-based assessments in the classroom is a rapidly growing movement within a larger movement of education reform. The use of these assessments calls for a clearer identification of what students should know and be able to do. In 1993 the National Education Goals Panel (1993) published *Promises to Keep: Creating High Standards for American Students.* This report outlines the importance of establishing two types of standards: content and performance. Content standards specify what students should know and be able to do. Performance standards specify the level to which the content standards should be mastered.

Norm-referenced tests provide some criterion-referenced information by indicating the number of questions answered correctly in a given content domain area; however, the norm-referenced tests used in the classroom today typically do not provide information as meaningful as that provided by standards-based assessments because the main purpose of the standardized, norm-referenced test is to compare students. For example, some norm-referenced tests have only three or four questions measuring a particular skill. In addition, norm-referenced tests often do not include questions that are very easy or very difficult because these questions are not typically useful in discriminating among students. The standards-based assessment usually focuses more upon ensuring that the difficulty level of the questions matches the specific learning targets as defined for the core curriculum (McMillan, 1997).

Understanding Performance Levels

In standards-based assessments, performance level descriptors are guidelines for determining what and how much students should know about a given content area at various stages of their formal schooling. Assessment training should include a discussion of performance levels, particularly the need to craft clear descriptors. Writing performance level descriptors requires careful analysis of the curriculum content standards in order to summarize dimensions of performance in a way that is clear and relevant to the standards. McMillan (1997) outlines six steps to follow when summarizing the dimensions of performance:

1. Identify dimensions of excellence.
2. Categorize and prioritize dimensions.

3. Clearly define each dimension.
4. Identify examples.
5. Describe performance continuums.
6. Try out and refine each dimension.

Stiggins (2001) stresses that assessment training must help educators or those developing the assessments to determine where and how evidence of academic proficiency will manifest itself. To identify performance criteria upon which to judge achievement, educators need to analyze the skills students are expected to demonstrate. This requires identifying the important elements that come together to make for sound performance.

Developing Content Domains and Understanding Validity

After determining the purpose of an assessment, educators must learn how to determine what should be assessed. Clear definitions of the purpose and content domains are important for all assessments, whether norm-referenced or criterion-referenced. Clearly defined content domains guide the entire assessment development process and aid in establishing the validity of the assessment. *Validity* refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 1999).

As integral parts of assessment training, establishing validity and determining content domains begin with a careful examination of the core skills and standards or learning targets to be measured by the assessment. For many standards-based assessments, the development of content domains first involves an in-depth analysis of the curriculum content standards for a given program. The purpose of this analysis is to give teachers, as well as other educators involved in this process, a full understanding of the fundamental principles underlying what should be taught. Knowing what to ask of and teach students is important because different achievement targets require the application of different assessment methods. In any assessment context, the assessment development process must begin with a clear vision of what it means to succeed in that context. Students are expected to know and understand specific subject matter, some of which they must know outright, and some of which they must be able to retrieve using references as necessary (Stiggins, 1999).

Well-crafted content domains or content domain specifications ensure that assessments measure what they are intended to measure. The following steps are recommended for training educators to develop

content domain specifications:

1. Review the purpose of the assessment.
2. Review the purpose for developing content domains.
3. Analyze the curriculum content standards for a given grade level and content area.
4. Determine what is to be assessed.
5. List the learning targets to be assessed.
6. Provide an indication of the relative importance of the content to be assessed.

Without a strong association between the assessment questions and the content domains, the questions will lack meaning and purpose (Osterlind, 1989). The following section illustrates the process used to develop a specific content domain specification.

Example of the Content Domain Specification Process

A group of teachers attend an assessment training workshop. They have reviewed the purpose of the assessment, and they understand the reason for developing content domains. The teachers have been asked to analyze the curriculum content standards for grade eight English language arts. After carefully analyzing the curriculum standards for their program, they determine that three content domains are represented by the curriculum content standards: using resources and following the research process steps, writing effective content and organizing clear paragraphs, and editing and revising paragraphs. The teachers write descriptions of each content domain. The following description is excerpted from one domain:

Subject area. Grade eight English language arts

Content description. Using resources and following the research process steps

Content domain description. Assessment questions in this domain will assess students' ability to understand and identify the steps in the research process, including identifying, collecting, and using sources of information.

Skills measured

Identifies and uses print sources to gather information
Identifies and uses technology to gather information
Identifies and uses media sources to gather

information
Identifies and uses the research process steps

Developing an Assessment Blueprint

After specifying the content domains, the next step in the training process is for educators to learn how to develop or review assessment blueprints. Typically the assessment blueprint will include a list of all standards to be assessed, organized by content domain. This blueprint outlines the number of assessment questions to be developed per learning target. In order for most standards-based assessments to be valid, there must be a close correspondence between the assessment content and the learning targets as specified in the curriculum content standards. Assessment blueprints further define what should be measured and what is important from the content domains. Good assessment blueprints produce reliable and valid assessments. As the *Standards for Educational and Psychological Testing* states:

> Important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. . . . Test developers often work from a specification of the content domain. The content specification carefully describes the content in detail, often with a classification of areas of content and types of questions. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct. (AERA, APA & NCME, 1999)

Instructional validity plays an important part in the development of assessment blueprints. *Instructional validity* measures the extent to which an assessment matches what is taught (McMillan, 1997). How closely does the test correspond to what has been covered or should be covered in the classroom? Have students had the opportunity to learn what is being assessed? The process of developing blueprints for various subject areas is often challenging because the nature of learning is qualitatively and quantitatively different across disciplines. For example, teaching and learning mathematics are often tied to a specified instructional sequence. Therefore, what students should know and be

able to do is often tied directly to the level and quality of their classroom instruction. Conversely, the learning targets for a grade eight English language arts standards-based assessment might be more generalized and not tied to an instructional sequence. As a result, assessment training must help educators learn how to use their professional judgment and knowledge of the curriculum to make decisions about the importance of different types of learning targets, the content to be assessed, and how much of the assessment should measure each target and content area. The assessment blueprint should provide for a wide coverage of the learning targets (Osterlind, 1989). The following steps are recommended for developing effective assessment blueprints:

1. Review the purpose of the assessment.
2. Review the purpose for developing assessment blueprints.
3. Analyze the curriculum content standards for a given grade level and content area.
4. Determine what is to be assessed and review the content domains.
5. List the learning targets to be assessed.
6. Provide an indication of the relative importance of the content to be assessed.
7. Determine the structure of the assessment, including recommended length, item difficulty, and higher-order thinking skills required.

Developing Assessment Question Specifications

Assessment training involves training educators to develop detailed specifications for writing assessment questions, often called *item specifications*. Item specifications ensure consistency throughout the entire assessment development process. Learning how to write and review item specifications is a crucial component of any comprehensive assessment training program. Through this training educators can begin to see the link between assessment and what is taught in the classroom.

Item specifications are one of the key requirements for a high-quality standards-based assessment. Although there are some similarities between the assessment blueprint and the item specifications, item specifications are usually more specific. They delineate the general characteristics of the questions for each curriculum content standard, and they provide information concerning the procedures for writing and reviewing test questions, including a detailed set of instructions from the assessment developer to the individuals writing the test questions so that the learning standards for a particular program can be

translated into good test objectives. For example, when developing an assessment blueprint for English language arts, one standard might read, "Applies standard rules of capitalization." What are the standard rules? Should all be tested? Should standard rules of capitalization be tested using multiple-choice questions or a writing performance assessment only? In developing item specifications or reviewing item specifications for "applies standard rules of capitalization," educators should bring to the task their knowledge of the curriculum and of students to determine exactly what capitalization rules should be included on the assessment at a given grade level and how these rules should be assessed.

Well-crafted item specifications must clearly define the purpose of the assessment and the content to be measured. The following steps are recommended for developing item specifications:

1. Review the purpose of the assessment.
2. Review the purpose for developing assessment blueprints and item specifications.
3. Review the analyses of the curriculum content standards for a given grade level and content area, including what is to be assessed.
4. Review the structure of the assessment.
5. Provide an indication of the relative importance of the content to be assessed.
6. Draft item specifications, including
   - a statement of the content domain
   - a statement of the curriculum standard to be measured
   - directions for how the test questions should be written
   - estimated difficulty of the assessment questions
   - guidelines for determining the cognitive level for learning targets
   - guidelines for how the answer choices should be written
   - any additional guidelines for how a particular learning target should be tested

Writing and Reviewing Assessment Questions

Educators regularly write assessment questions; however, most educators have not received training in how to write or review good test questions. One goal of assessment training is to equip educators with the tools necessary to write and review good test questions. A model training session should include an overview of the basic item writing guidelines, as well as any specific guidelines for a particular program. The assessment blueprint, as well as the item specifications,

should also be discussed, along with any general information about basic item writing principles. Educators should then be given the opportunity to write items to measure specific content standards. The training should also include a review of the items written, with suggestions for revision. The following steps are recommended steps for reviewing and writing assessment questions:

1. Review the purpose of the assessment.
2. Review the content domains, the assessment blueprint, and the item specifications.
3. Review guidelines for writing and reviewing good, reliable, and fair test questions.
4. Draft and review items.

Although there are many well-established guidelines for the writing of good, reliable, and fair test questions, the following section focuses on some of the most important guidelines to include in assessment training for educators.

Match the Question to the Content Standard

The first criterion for writing good test questions is that there must be a high degree of congruence between a particular question and the key objective of the test (Osterlind, 1989). How well does the question match its intended objective? The congruence criterion is the assessment writer's or reviewer's primary consideration because it is at the heart of validity. Careful attention to the educational significance of the assessment questions ensures that the assessment mirrors sound instructional practices. Training should include helping educators write questions that measure the specific content objectives. Every question has a purpose, and this purpose should be clearly defined in the question specifications. The standard and the benchmark help to define the nature of the test question.

A major prerequisite for writing and reviewing questions that match the standard is strong knowledge of the content area to be probed by the questions. As an integral part of the training process, teachers should clearly establish the close correspondence between the curriculum content standards and the test questions. The content of each test question must be crafted carefully so that the question measures what is important and what can be successfully taught and learned in the classroom.

Knowledge Specificity

Educators must carefully consider the educational significance of each assessment question. Assessment training should therefore include a discussion of the concept of knowledge specificity. *Knowledge specificity* refers to a continuum of overly specific to overly general questions. Most questions should be written with this continuum in mind (Haladyna, 1999). Questions should not measure simple recall of facts or be classified as "so what" questions. The "so what" question is at the lowest level on the continuum and usually asks for knowledge that has little or no value for assessing a student's progression toward mastery of the learning targets (Frary, 1995).

Item Quality

Potential problems with a test include poorly worded questions, reading or writing demands that require more than a mastery of the material being tested; questions with more than one correct response; incorrect scoring; or racial, ethnic, or gender bias. In addition, the student may experience extreme test anxiety or interpret test questions differently from the author's intent, as well as cheat, guess, or lack motivation. Further, the assessment environment could be uncomfortable, poorly lighted, noisy, or otherwise distracting (Stiggins, 1999). Any of these situations could give rise to inaccurate test results. To prevent problems related to test questions, assessment questions should be well written, following a uniform style. Although there are several published guidelines for writing and reviewing assessment questions, the following list summarizes the major considerations for writing good, reliable, fair test questions.

A good question

- has one and only one clearly correct answer
- is structured around one main idea or problem
- measures the objective or curriculum content standard it is designed to measure
- is at the appropriate level of difficulty
- is simple, direct, and free of ambiguity
- makes use of vocabulary and sentence structure that are appropriate to the grade level of the students being tested
- contains answer choices that that are plausible and reasonable in terms of the requirements of the question, as well as the student's level of knowledge

- contains answer choices that are parallel in grammatical structure and content
- contains answer choices that relate to the question
- reflects good and current teaching and learning practices in the subject area
- is free of bias

A bad question

- provides clues (within a question or within a test form)
- is considered a "trick" or "cute" question
- contains an answer choice that would eliminate another answer choice
- contains vocabulary and idiomatic phrases that could be unfamiliar to students
- asks about trivial information

Levels of Thinking

An important objective of classroom instruction is to help students acquire and use higher-order thinking skills. Assessments, therefore, must include questions that require higher-order as well as lower-order thinking, and educators should be trained to evaluate each question they write in terms of the levels of thinking required to answer the question.

Defining and measuring the levels of higher-order thinking has been a major challenge to educators for many years. Each question in a test measures a specific behavior, and students may respond to questions with a pattern of right and wrong answers, but no one really knows the exact mental processes used in making the correct choices on a test. For any test question, the test taker may appear to be thinking at a higher level, but in actuality, he or she may be remembering identical statements or ideas previously presented. A group of educators may agree that a given question appears to measure one type of behavior, when in fact it may measure an entirely different type of behavior simply because each test taker brings a unique set of experiences to the test (Haladyna, 1999).

Perhaps the best-known source for learning targets and cognitive processing is the *Taxonomy of Educational Objectives* (Bloom, 1956). *Bloom's taxonomy* is probably the most widely used scheme for labeling levels of cognitive processes. Using this taxonomy, assessment questions can be classified into one of three cognition levels: recall, application, and analysis. *Recall* questions are written to measure students' ability

to remember isolated facts, concepts, principles, processes, procedures, or theories. When students respond to these questions, the primary cognitive function they use is memory. *Application* questions are written to measure students' ability to provide simple interpretations or limited applications of data or information. Questions written at this level typically require some problem-solving skills. *Analysis* questions are written primarily to measure students' skills in evaluating data and problem solving. Responding to these questions involves application of good judgment and problem-solving skills. Analysis questions involve higher cognitive processes than do the other types of questions (Vacc, Loesch, & Lubik, 2001).

Bloom's taxonomy was developed many years ago, and many educators believe that the taxonomy is no longer adequate for defining levels of cognitive processing. In fact, since the development of Bloom's taxonomy, there have been many changes in the educational and psychological theories that formed the basis for the taxonomy.

Current theories emphasize thinking processes, characterize the learner as an active information processor, and stress domain-specific thinking and learning (McMillan, 1997). For example, the *dimensions of learning* is an instructional framework based on current research and learning theory. Initially, the dimensions of learning framework was designed to help educators plan curriculum and instruction more effectively by using what is known about how students learn. The framework's strong grounding in research and theory, however, makes it a natural partner for assessment (Marzano, Pickering, & McTighe, 1993). Following are the five dimensions of learning:

1. Maintaining positive attitudes and perceptions about learning
2. Acquiring and integrating knowledge
3. Extending and refining knowledge
4. Using knowledge meaningfully

Dimension 5: Developing productive habits of mind

The five dimensions of learning can be used to address current content standards, including acquisition and integration of knowledge, complex thinking standards, and reasoning processes standards (Marzano, Pickering, & McTighe, 1993). Whether or not educators use Bloom's taxonomy or the five dimensions of learning, they must be trained in understanding the precision of the cognitive level definitions adopted for use, and they must be trained to consider

carefully the precision with which particular test questions may tap specific levels of mental processing (Osterlind, 1989)

Item Difficulty

Most standards-based assessments used in the classroom today should include test questions that have a range of difficulty so that all students can demonstrate what they know and are able to do. The level of thinking required to answer a particular question is not the same as the difficulty of the question. For example, a question calling for a student to analyze an easy reading passage may be a much easier question than a question that asks a student to demonstrate comprehension of a difficult reading passage. Therefore, a major component of assessment training is to help educators understand item difficulty and that determining the difficulty of a question requires teacher judgment.

Summary

It has often been said that there is a gap between assessment and instruction in the classroom. Often the instruction in the classroom is not geared toward the same objectives as those measured on the assessment, or the assessment may, in fact, fail to provide information about students' strengths and weaknesses as real targets for further instruction. The assessment training model presented in this chapter may serve as a starting point for providing educators with the tools necessary to make the critical connection between instruction and assessment. Through in-depth knowledge of the purpose of the assessment and the assessment development process, instruction can be placed on the same educational continuum as the standards-based assessment.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Frary, R. B. (1995). More multiple-choice question writing do's and don'ts. *Practical Assessment, Research and Evaluation. 4*(11), 1–4.

Haladyna, T. (1999). *Developing and validating multiple-choice test questions.* Mahwah, NJ: Lawrence Erlbaum.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Aurora, CO: McRel Institute.

McMillan, J. H. (1997). *Classroom assessment: Principles and practice for effective instruction.* Needham Heights, MA: Allyn and Bacon.

Mehrens, W. A. (2000). Selecting a career assessment instrument. In J. T. Kapes & E. A. Whitfield (Eds.), *A counselor's guide to career assessment instruments* (4th ed.). Alexandria, VA: National Career Development Association.

National Education Goals Panel. (1993, November). *Promises to keep: Creating high standards for American students.* (Report on the review of education standards from the Goals 3 and 4 Technical Planning Group to the National Education Goals Panel.) Washington, DC: Author.

Osterlind, S. J. (1989). *Constructing test questions.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stiggins, R. (1999). Are you assessment literate? *High School Magazine, 6,* 2–6.

Stiggins, R. (2001). Sound performance assessments in the guidance context. In G. R. Walz & J. C. Bleuer (Eds.), *Assessment issues and challenges for the millennium.* Greensboro, NC: CAPS Publications.

Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). Writing multiple-choice test questions. In G. R. Walz & J. C. Bleuer (Eds.), *Assessment issues and challenges for the millennium.* Greensboro, NC: CAPS Publications.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").